Chemometrics and Intelligent Laboratory Systems, 11 (1991) 161–174
Elsevier Science Publishers B.V., Amsterdam

RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation

Part 3: The evaluation

J.A. van Leeuwen *, L.M.C. Buydens, B.G.M. Vandeginste and G. Kateman

*Department of Analytical Chemistry, Catholic University Nijmegen, Toernooiveld 1, 6525 ED Nijmegen
(Netherlands)*

A. Cleland and M. Mulholland

Philips Scientific, York Street, CB1 2PX Cambridge (UK)

C. Jansen and F.A. Maris

Organon International B.V., P.O. Box 20, 5340 BH Oss (Netherlands)

P.H. Hoogkamer and J.H.M. van den Berg

*Duphar B.V., Research Laboratories, Analytical Development Department, P.O. Box 900, 1380 DA Weesp
(Netherlands)*

(Received 15 June 1990; accepted 15 February 1991)

Abstract

Van Leeuwen, J.A., Buydens, L.M.C., Vandeginste, B.G.M., Kateman, G., Cleland, A., Mulholland, M., Jansen, C., Maris, F.A., Hoogkamer, P.H. and Van den Berg, J.H.M., 1991. RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 3: The evaluation. *Chemometrics and Intelligent Laboratory Systems*, 11: 161–174.

In order to establish the confidence of target users in an expert system, it is important that it has been submitted to a thorough testing procedure. In the case of RES, an expert system on ruggedness testing in HPLC, the testing procedure is divided into two parts: a validation and an evaluation part. In the validation part, the performance of the expert system is measured against the expert's performance using simulated test cases. In the evaluation part, the suitability of RES is assessed by evaluators in a real laboratory environment. RES has been evaluated in two laboratories. In one laboratory the emphasis of the evaluation was on the knowledge in RES. In the second evaluation the usability of RES with respect to users who were inexperienced in formalized testing was emphasized. The results and comments of the evaluators are discussed. RES has been found to be useful in practice.

INTRODUCTION

The application of expert system technology to chemical analysis is investigated in the Esprit project 1570 ESCA (Expert Systems in Chemical Analysis). Expert systems are computer programs that contain the knowledge of a recognized expert in a certain knowledge domain. This expert knowledge is incorporated into the expert system using a form of knowledge representation e.g. IF...THEN...rules [1]. The aim of an expert system is to bring an expert's knowledge and experience to greater use by making it available to a greater public in the form of a computer program.

The well defined domain of method development in HPLC was selected as the knowledge domain in the ESCA project. In this knowledge domain, the ESCA project has produced a number of expert systems [2]. The main knowledge domains are: initial guess of HPLC method conditions [3], selection of an appropriate criterion for the optimization of the mobile phase [4], optimization of the instrumental parameters and operating conditions [5,6] and method validation [7–11].

In the development of these expert systems, three clearly defined phases have been defined [12]. The first phase consists of selecting appropriate expert system development tools that meet the requirements of the knowledge domain [13]. In the second phase the tools are used to build the expert systems. In the third phase the expert systems are investigated in terms of their performance and use in the analytical laboratory.

The third phase, the testing phase, is an important phase in expert system development. The practical use of the expert system largely depends on the successful conclusion of a thorough testing procedure. Many expert systems never reach the stage of practical use because the test procedures applied did not establish enough confidence in them [14]. Apart from real problems with the performance of the expert system, the lack of acceptable expert system test procedures also causes a lack of confidence in expert systems.

In the ESCA project, a strategy has been developed for the expert system testing phase. It is intended to establish sufficient confidence in the

TABLE 1

Definition of validation and evaluation

Validation: the test on whether the system responds to the requirements specified by the developer.

Evaluation: the test on whether the system corresponds to the expectations and requirements of the users of the system.

HPLC method validation: test procedures performed on an HPLC method to test its performance with respect to features such as specificity, precision, sensitivity and method limitations.

expert systems developed in ESCA by combining theoretical testing of the system using simulated test-cases with field tests in analytical laboratories. In this paper, the test strategy developed in ESCA is described, and the results of the Ruggedness Expert System, RES, in this test procedure are described [15,16].

The knowledge domain of RES, the ruggedness test, is part of an extensive experimental scheme, normally performed on a new HPLC method. This phase is commonly referred to as HPLC method validation. The concept of validation is also used in the expert system testing procedures, but here it has a different meaning: in expert system testing, validation is used for the test procedures that compare the performance of the expert system against the expert. Table 1 presents descriptions of HPLC method validation and expert system validation. In this paper, HPLC method validation will be referred to as method validation and expert system validation will be referred to as validation or validation of the expert system.

RES

Method validation plays an important role in HPLC method development. In method validation, a newly developed HPLC method is submitted to a number of tests in order to establish its performance. A method validation procedure usually includes tests on accuracy, precision, sensitivity and specificity [17–19].

In method validation, the HPLC method can be submitted to a ruggedness test [20–22]. In a

ruggedness test the performance of the HPLC method is tested under slightly different conditions than the normal operating conditions. The purpose of the test is to simulate changes in the operating conditions of the method when the method is transported from one laboratory to another. Examples of conditions that will undoubtedly vary from laboratory to laboratory are temperature, batch of column material, etc. In the ruggedness test the effect of these variations on the performance of the method can be established by simulating them in the laboratory in which the method was developed.

A ruggedness test consists of a number of steps [15]. The most important steps are the following.

- Selection of factors to test. There is a large number of factors that could be tested in a ruggedness test because they can influence method performance. However, it is impossible to test them all. Therefore, a selection must be made of ‘most suspect’ factors that will be tested in the ruggedness test. Making this selection requires a great deal of experience in HPLC ruggedness testing and method development.
- Selection of the experimental design to test the factors. The number of experiments in the ruggedness test must be kept as low as possible. Numerous experimental designs exist that could be used in a ruggedness test. The most appropriate one must be selected from this collection of designs, taking account of the number of factors to be tested.
- Performance of the experiments. The experimental work of measuring the performance of the method under the various circumstances.
- Calculation of the statistical parameters from the experimental data. Parameters such as main effects are calculated. The main effects indicate the influence of a certain factor on the performance of the method.
- Deduction of chemically relevant information from the statistical parameters. The statistical parameters are translated into chemically relevant information. For instance, if the main effect for resolution on the factor temperature is larger than 10%, a certain action must be undertaken: the initial resolution of the method must be increased, for example.

- Advice on improvements to the method, such as the initial resolution required to resolve the resolution problem.

Some of these steps are typically heuristic (based on experience) and are therefore suitable for implementation in an expert system, e.g. the step concerning the selection of factors to test [9]. Other steps are more of an algorithmic nature and are therefore more suitable for implementation in an algorithmic language, e.g. the calculation of the statistical parameters. However, a software system that must guide a user through a ruggedness test must contain heuristics as well as algorithms. RES guides a user through a complete ruggedness test. RES has some features that allow the integration of heuristic and algorithmic knowledge. An important feature is the representation of the knowledge in modules. These modules correspond to the steps in the ruggedness test. Each module is either heuristic or algorithmic.

EXPERT SYSTEM TEST PROCEDURES

Testing expert systems differs from testing algorithmic software in a number of ways. An important difference lies in the impossibility of assessing all the program code and all possible paths through the code in an expert system testing procedure. Because an expert system usually consists of rules and procedures that chain into each other, the number of possible pathways through the code is very large. It is therefore important to establish in a test procedure what the limits of the system are. The risk that a future user enters information into the expert system that the system cannot recognize should be reduced as far as possible during the testing procedure.

Another difference from conventional software is that, in many cases, it is impossible to define what a ‘good’ result of the expert system is. As the expert system contains the knowledge of an expert, it contains certain assumptions that the expert has made. These assumptions will be reflected in the results produced by the system. The expert system can also contain areas of knowledge that the expert is not completely sure of. Such areas of knowledge can result in less accurate results, just

as the expert will solve some cases better than others. Because of the impossibility of ensuring that the expert system always produces the best possible answer, the user must have ample possibilities to interact with the system and change the system's intermediate results during a consultation.

Problems of this type are not addressed in algorithmic software testing procedures, which usually involves the identification of programming errors by comparing the solution of the program with a calculated solution. However, they do not take into account any uncertainty in the solution, nor do they take account of the fact that, in an expert system, a large part of the code will not be used in a normal consultation. Expert systems therefore need special testing procedures.

Due to the nature of expert systems, it is to be expected that general testing procedures are hard to define. Expert systems differ very much in their knowledge bases and it is especially this part that needs thorough testing. The few cases of extensive testing described in the literature are dedicated to a specific expert system. No general strategy exists that gives more detail than that the expert system should pass a number of phases before it can be evaluated in practice [14,23,24]. As a result, no general guidelines like the IEEE standards for conventional software testing exist for expert system testing [25], except for exclusively rule-based systems. For such systems, a test procedure can be constructed that directly tests for such failures as conflict, redundancy and circular reasoning [26]. Rule-based systems with a fixed rule syntax may even include the possibility of automatic adaptation [27].

The ESCA approach

The objective of the test strategy developed in ESCA is to assess the performance of the expert system in comparison with the expert's performance. If it can be shown that the expert system performs as well as the expert this will establish the confidence of future users in the expert system. A second objective is to assess the usability of the expert system in real laboratory situations.

Both objectives can be achieved by comparing the performance of the expert systems with the performance of experts when they are confronted with the same problems. This requires that the same problems are presented to both the expert system and the expert, and that the answers are verified in practice, if possible. In many cases it is not advisable to organize a comparison between a real expert and an expert system. An important disadvantage of such a comparison is the large number of test cases needed to be able to decide on the ranking of the expert vis à vis the expert system. Performing a large number of practical tests may become costly and may extend over several weeks and involve expensive equipment and labour.

The strategy developed in ESCA is based upon the consideration that only the expert can decide whether the expert system performs as it should perform. The expert can judge whether the system does contain his knowledge and whether the knowledge is used correctly by the system. The performance of the system can be measured by comparing the system's results with the expert's solutions using simulated test cases. The test cases can be simulated to contain various levels of complexity. This part of the test procedure can be seen as a theoretical test and is called validation (Table 1). The practical side of the test strategy is concerned with the question of whether the expert system is useful in a laboratory environment. Only future users can decide upon this issue. Users will decide whether the expert system is useful, easy to use, etc. They will also be able to assess the benefit they get from using the system. This part of the testing procedure, the evaluation, must be performed in a real laboratory situation, preferably with practical laboratory work being done to verify the results of the system. Involving future users in the evaluation procedure also makes it possible to gain information on how users feel about getting expert advice from software systems. The division of the testing procedure into a validation part and an evaluation part has the advantage that simulated test cases can be used in the validation part. It reduces the cost of the testing procedure, not only by reducing the practical work, but also by enabling the expert to work with batches of test cases.

Test criteria

Because of the difficulty in defining what a 'good' result of the expert system is, it is difficult to define what criteria should be used to test an expert system. In the validation, it is possible to set a certain percentage of test cases on which the expert and the expert system must agree before the expert system can be evaluated. In the evaluation however, it is more difficult to set criteria. The knowledge incorporated in the expert system must be valid. Also, the expert system must be understood by a user and must allow the user to make optimum use of the knowledge in the expert system. These two features interact with each other during the evaluation. The evaluators only have a limited insight into the contents and organization of the knowledge base, so it is difficult for them to identify the cause of any problems they may encounter. Nevertheless, it is possible to identify a number of criteria that are important in the day-to-day use of the expert system. Some of these criteria can be tested in the validation. Others can only be tested in the evaluation (Table 2). The list of criteria can be extended by a number of related criteria. However, it is difficult to make an extensive list with all criteria that should be tested in an evaluation procedure. The evaluators were therefore asked to pay attention to the criteria, but they were also asked to make comments which were not related to a specific criterion.

Accuracy

The accuracy is a measure of how good the expert system's solution resembles the expert's solutions. The accuracy may be expressed as a

percentage of solutions that are acceptable to the expert. A second aspect of accuracy is the repeatability of the system's solution. If a case is resubmitted to the system, the same answer should be obtained. Because the pathways through the knowledge base of the system are not predefined, a test should be performed of whether the system takes the optimum pathway each time. One way of testing this is by repeated consultation of the system on the same test case.

Completeness

The completeness of the expert system can be seen from two different points of view, the software point of view and the user point of view. From the software point of view, the expert system is complete if it does not allow input that is not covered by the knowledge base. It should also not end any consultation without producing an output. In practice, it is difficult to test all possible pathways. By testing the expert system with a wide range of test cases it can be shown that the system is complete to a large degree. Such testing can best be performed in the validation phase. In the evaluation, the completeness of the knowledge base can be established with respect to the requirements of day-to-day users. The knowledge base is complete if the evaluators do not discover any missing parts that are essential for proper use of the expert system.

Usability

It should be possible for users to operate the expert system with a minimum of supervision and training. The usability depends on the user interface and the explanation facilities of the expert system. Users should be able to interact with the system not only to get advice, but also to change any intermediate results in the system to their own preference. The usability also includes an assessment of the usefulness of the knowledge base contents.

Consistency

It is important that the various modules in the system communicate with each other in a consistent way. The user must be able to change intermediate results to reflect his own ideas. It

TABLE 2
The criteria to evaluate

	Validation	Evaluation
Accuracy	x	
Consistency	x	x
Completeness		
software	x	
user		x
Usability		x
Quality of advice		x

should be impossible for the user to make changes that lead to system crashes or invalid results.

Quality of advice

It is often difficult to identify correct and incorrect answers. Instead, there is a large gray area between correct and incorrect. Therefore, the expert system should at least produce solutions that are acceptable to the user. The quality of advice can be measured as the changes that the user will make to solutions produced by the expert system.

The validation of RES

Validation involves the selection of a number of test cases which demonstrate a variety of situations. The expert predicts the answers, and then tests the problems on the expert system. If there are any inconsistencies, the expert should decide from one of the following situations:

- some of the knowledge in the system is incorrect;
- some knowledge is missing from the system, which should be added to the system;
- the knowledge in the system is deemed to be satisfactory, and the system is regarded as providing good answers to the problems set. In this case the expert made a mistake and is corrected by the system.

Validation shows whether the systems are providing good expert answers to problems falling within the intended scope of the systems. The problems that are selected will in some way be biased as they were chosen by the experts themselves. If there is a bias, it will therefore be a bias towards a realistic range of problems.

In order to prevent the introduction of problems when the expert decides to fix a problem or add new knowledge to the system, a regression test is used. It is used to perform quick tests on versions of the expert system to ensure that the basic functionality has not changed in an unexpected way. The intention is to define a range of realistic problems which will be submitted to the expert system, every time changes are made to it, to ensure that it continues to give valid results.

The validation process performed on RES con-

sisted of a number of simulated test cases that were submitted to the system. As the system has a modular structure, it was decided to validate each module separately. Because the modules differ in nature from algorithmic to heuristic, different testing procedures were used for the different modules [16].

The algorithmic modules were validated using simulated test cases with pre-calculated results. The approach used in their validation was similar to approaches proposed for conventional software testing, such as the IEEE standards [25]. Such conventional software testing standards include bug reports, identification of responsibilities and logging of bug fixes. The algorithmic modules in RES were tested with approximately 50 simulated test cases to ensure their bug-free operation. After being tested with the simulated test cases, two real data sets were submitted to the algorithmic modules as an ultimate test. The heuristic modules were also validated using simulated test cases. A set of 11 test cases was generated to cover as much of the knowledge in the system as possible. The coverage of the test cases over the knowledge added up to some 30%.

In the validation of the heuristic modules, the concept of regression testing was used. However, to decide which test cases should be part of the regression test set was not straightforward. In fact, it was only for the rule-based modules that it was possible to set criteria on which test cases to include in a regression test. For instance, for the factor choice module the regression test set was selected from the eleven test cases available in the validation. On practical considerations, the number of test cases in the regression test set was limited to four. For every test case a rule trace was made which showed which rules were triggered. Thus, the number of rules used in each test case was assessed. The test cases were ranked according to the number of rules they triggered. Also, the rules triggered in only one test case, the 'single' rules, were listed. The two test cases with the largest number of 'single' rules, rules that were triggered only in that test case, were included in the regression test set, together with the two test cases that triggered the largest number of rules overall. With this regression test set, the heuristic

modules were refined, leading to a success rate of 80–90%.

The evaluation of RES

Because of the limited resources available and the time needed to perform a ruggedness test (Table 3), the number of evaluators had to be limited to two. The first evaluator was asked to concentrate on the contents of the knowledge base of RES. Evaluator 1 has a background in HPLC method validation and formalized ruggedness testing and had already performed a number of ruggedness tests before evaluating RES. The second evaluator also has a background in HPLC method validation. However, this evaluator has little experience in systematic ruggedness testing using experimental designs, but is experienced in ruggedness testing as part of method development and method optimization using univariate methods. Therefore Evaluator 2 was asked to concentrate on such aspects of RES as the user interface and usability.

Evaluator 1 must be regarded as an expert user of RES. For this evaluator RES should be an advisory system that the expert user can use as a check on his own conclusions. Evaluator 1 can assess the validity of the knowledge and the reasoning strategy in RES. Because Evaluator 1 already has a well established ruggedness testing strategy, some conflicts between the strategy embedded in RES and the strategy used by Evaluator 1 are to be expected. Evaluator 2 represents a typical user of RES who uses RES as a guide while performing a ruggedness test.

Three different types of test cases have been used for the evaluations, simulated test cases, historic data of ruggedness tests and ruggedness tests performed following the recommendations of the expert system. Some of the test cases performed by Evaluator 1 are simulated test cases similar to the test cases that have been used in the validation of RES. The evaluators selected the test cases from a list of methods that will undergo a ruggedness test in their laboratory in the near future. As these test cases are related to practical test cases, any bias in the selection of the test cases will be a bias towards reality.

A number of real test cases will also be submitted to the system. The first evaluator submitted two of his previously performed ruggedness tests to RES. Comparison of the results of the system with the original results gives an estimate of the correctness of the knowledge base. Also, differences in strategy between the expert system and the evaluator will be identified quickly. The second evaluator performed two ruggedness tests guided by the expert system. This evaluation primarily resulted in conclusions about the usability of both system and ruggedness test and about man–machine interface issues.

Results

In general the evaluators found the system to be useful. The modules contain knowledge that is valuable in a laboratory and the system contains parts, such as the factor choice module, that contain knowledge that cannot be found elsewhere. The system contains a consistent line of reasoning and produces results that can be used in method validation and method improvement. The evaluators also made a number of comments on areas where they thought the system could be improved. These comments on the system can be divided into three categories that represent possible sources of dissatisfaction with the system. The first source of problems is the discovery of errors that should have been eliminated in the validation phase, the reasoning errors, programming errors and unclear areas in the system. Although the validation was as careful and complete as possible, it is inevitable that some errors persist and remain to be dis-

TABLE 3

Time involved in performing a ruggedness test with the help of the expert system

Step	Days
1. Input the method, select factors, edit factors and select design	1–1.5
2. Carry out the experiments	5–9
3. Create the data file	1–2
4. Input of the data file into the expert system and report interpretation	1–2

covered in a later phase. Fortunately, the occurrence of such problems was limited. If possible they were repaired immediately to prevent them from interfering with the evaluation. For instance, a problem with the calculation of standard errors revealed by Evaluator 1 was corrected in the version submitted to Evaluator 2.

A second source of problems is the possible difference in ruggedness testing strategy that the evaluators may already have developed. The expert system contains the knowledge of a recognized expert on ruggedness testing. However, this does not mean that the strategy implemented in the expert system is the only useful strategy. Other experts are supposed to follow strategies that are broadly similar, but differences of minor, and sometimes major, importance can be found. This is not a real problem as long as the strategy employed by the expert system is clear and correct. Users of the expert system should be able to identify differences between their strategy and the expert system's strategy quickly and at an early stage.

Thirdly, the evaluators proposed a number of extensions to RES that they would find useful. Suggestions ranged from the method description, where more input variables should be allowed, to the interpretation of results, where additional statistical tests could be useful.

Evaluator 1

Evaluator 1 concentrated on the contents of the knowledge base. The evaluator tried to answer the following questions in the evaluation process:

- are all well known factors present in the system?
(completeness)
- is it possible to edit the experimental design?
(usability)
- is the statistical analysis valid?
(consistency)
- are the conclusions practical?
(quality of advice)

The evaluation procedure consisted of five test cases. Two of these are ruggedness tests that had

been performed at the evaluator's laboratory at an earlier stage. The method description and the experimental results are entered into the system to check whether RES and the expert agree on the outcomes of the ruggedness test. The three other test cases are ruggedness tests that are currently in course of preparation. The consultation of RES for these test cases has not been taken further than the selection of the design stage.

Comments

One of the crucial points in the evaluation of RES is whether it includes the majority of factors normally tested in a ruggedness test. RES cannot advise on factors that are not in the knowledge base so the omission of important factors would be a severe problem. The expert system appeared to be acceptably complete (Table 4). One of the factors, the sample weight, is tested differently by evaluator and expert system. The expert system advises a variation of the sample weight, whereas the evaluator normally varies the injection volume. Both approaches are basically the same and can be transformed into each other by simple calculations. Only four factors were identified that were not present in the expert system. From these four, one is related to a specific composition of the solvent of the sample that differs from the mobile phase composition (water in sample) and the others are related to factors already present in the expert system. The plate-number change is related to the column batch factor. The testing of a second additive (additive-2) is related to the testing of the first additive. Also, the testing of the ratio between solvent 2 and 3 is related to testing the percentage of solvent 1. Adding these factors to the system must be discussed, but it is relatively easy.

With respect to the first category of problems, unclarities and errors in the system, Evaluator 1 misinterpreted a major input item that influenced part of the evaluation process. As Evaluator 2 initially made the same misinterpretation, it is clear that the user interface of RES can be improved with respect to the messages displayed. This misinterpretation had the largest impact on the selection of factors (Table 4). In a number of the test cases a better agreement between the

TABLE 4

Factor selection results for Evaluator 1 and 2

Possible factors in the expert system	Results of the test cases on factor selection	
A sample weight	Sample 1.1 (Evaluator 1)	
B shake time	factors selected by evaluator:	a, o, q, s, x, 1, 4
C sonicate time	factors selected by expert system:	A, C, M, T, X
D heat temperature	Number of same factors:	2
E pore size 1	Sample 1.2 (Evaluator 1)	
F pore size 2	factors selected by evaluator:	a, o, q, x, 1, 2
G wash volume	factors selected by expert system:	A, M, P, T, X, Y
H extraction volume	Number of same factors:	2
I extraction 1	Sample 1.3 (Evaluator 1)	
J extraction 2	factors selected by evaluator:	a, n, o, p, q, r, u, x
K centrifuge minutes	factors selected by expert system:	M, P, X
L dilution	Number of same factors:	2
M data-handling	Sample 1.4 (Evaluator 1)	
N pH	factors selected by evaluator:	a, b, h, n, o, q, r, u, x
O temperature	factors selected by expert system:	H, M, P, X
P buffer concentration	Number of same factors:	2
Q solvent%	Sample 1.5 (Evaluator 1)	
R additive concentration	factors selected by evaluator:	a, n, o, p, q, r, u, x, 3, 4
S flow rate	factors selected by expert system:	A, M, P, X, Y
T manufacturer	Number of same factors:	3
U batch	Sample 2.1 (Evaluator 2)	
V r_f -range	factors selected by evaluator:	b, c, m, q, u, x, y
W filter	factors selected by expert system:	K, M, T, X, Y
X wavelength	Number of same factors:	3
Y UV time constant	Sample 2.2 (Evaluator 2)	
Z r_f time constant	factors selected by evaluator:	m, q, r, u, x, y
Factors added by the evaluators:	factors selected by expert system:	M, Q, R, T, X, Y
1 water in sample	Number of same factors:	5
2 other plate-number	Sample 2.2 (Evaluator 2)	
3 additive-2	factors selected by evaluator:	m, q, r, u, x, y
4 ratio solvent-2,3	factors selected by expert system:	M, Q, R, T, X, Y
Sample preparation factors:	Number of same factors:	5
Data-handling factor:	A-L + 1	
Chromatograph factors:	M	
Column factors:	N-S + 2-4	
Detector factors:	T-U	
	V-Z	

factors chosen by the expert and the expert system would have been found if the HPLC method had been described correctly. The mistake prevented any solvent-related factors from being selected, so the solvent percentage and additive concentration factors were never selected by the expert system. They would have been selected in the majority of cases if the mistake had not been made. However,

the evaluator still felt that despite the differences between his results and the system's results, the factor selection part is of great value to an unexperienced evaluator. Sample preparation factors are found to be especially difficult to predict.

The selection of the experimental design showed a problem of the second category, differences in strategy between the expert system and the evalua-

tor. The design used by the evaluator is a fractional factorial design (2^{k-p}) centered at the nominal level. This design is not included in the set of designs supported by the system. RES supports (reflected) fractional factorial designs. The main difference in these designs is the number of experiments required. The evaluator's approach requires fewer experiments but is best suited for the testing of stochastic factors that have a linear response curve in the range covered by the ruggedness test. Non-linear and deterministic factors can be incorporated into the design at the expense of an additional factor to be tested. This requires additional experiments so, depending on the number of non-linear and deterministic factors, the two strategies can resemble each other in the number of experiments.

Only two test cases were available for the evaluation of the calculation module and the diagnosis module. A difference in approach was found in the calculation of statistical results. The evaluator used a different outlier test than RES. Also, the decision on whether a main effect is significant or not was treated differently. The evaluator related this decision to the reproducibility of the method found in previous tests (see also the Discussion below). This is not possible in RES, where main effects are compared with predefined limits.

In general, the main comment is that the system is more suitable for an inexperienced user than for an experienced user. In this case, in particular, where the user already has a well

established ruggedness testing procedure, the system is not flexible enough at some points to accommodate a procedure other than the one embedded in the system, so the usability of RES in its present form is somewhat limited for this evaluator (Table 5).

Because of the error at the input of the HPLC method, it was difficult to estimate the quality of advice on the basis of the number of changes made to the system's advice by the evaluator. Still, the evaluator had the feeling that the quality of advice was acceptable.

Evaluator 2

The second evaluation took place at the Quality Assurance Department of Evaluator 2. At this department there was little experience with formalized ruggedness testing using experimental designs but ample expertise in method validation and great interest in learning more about ruggedness testing. The evaluation concentrated on four main questions regarding input and output of the system, since it was difficult for the evaluators to evaluate the contents of the knowledge bases:

- is the input complete and clear? (completeness/usability)
- agreement with the proposed experimental plan? (quality of advice)
- are the experimental results processed correctly by the expert system? (consistency)
- does the expert system arrive at a correct and useful result? (usability)

TABLE 5

Performance of RES with respect to the criteria

	Evaluator 1	Evaluator 2	Validation
Accuracy			+
repeatability			+
Completeness			
software			+
knowledge base	+	+	
Consistency	+	+	+
Usability			
user-interface	–	–	
flexibility	–	+ / –	
explanation		+ / –	
Quality of advice		+	

The evaluation consists of two test cases that are carried through the expert system completely, including the experimental work. The two test cases were selected to emphasize different parts of the expert system. In the first test case, the sample pretreatment is rather complex. Three different solvents are added. Sonicating, shaking and centrifugation are all used for extraction. Although two different detectors are used for the two components, the chromatography is relatively simple. No critical resolutions will occur. The resolution between the components and the internal standard are both about 10.

In the second test case, the separation is the most critical part. In one run six components are determined. The resolution between two of the six components is sensitive to mobile phase changes. Sample pretreatment is very simple. Besides the three main components, three decomposition products are determined in the same run.

Comments

There was a problem getting the correct hardware installed. A workable solution was found by a partial re-implementation of the expert system, deleting some less essential parts. These parts, mainly the explanation facility, could be consulted on the original hardware in the course of the evaluation.

In the first test case, some problems appeared with the input of the HPLC method into RES. A number of features of the method could not be entered. The problems with the input of the methods are:

- sonicate and shake could not be selected together;
- a fluorescence detector was missing;
- it was impossible to select more than one detector, only the UV variable could be entered.

Furthermore, the same interpretation error of one of the input questions was made as by the first evaluator. This problem was identified at an early stage and the evaluation process was not influenced by it.

In the first test case, the number of factors was changed by the evaluator (Table 4). However, most of the changes could easily be explained. The solvent factor was added because, in this method, the mobile phase is premixed. If water and methanol are mixed by the HPLC equipment, an unstable baseline can occur due to the difference in absorption between water and methanol at 205 nm. Therefore, only one solvent line was selected. It appeared that the solvent was not selected as a factor if only one solvent line is selected. If two solvent lines are selected, solvent was selected at 17% and 23% (nominal level 20%). The evaluator changed these levels to 18% and 22% because they seemed to be rather large for this method.

The 'manufacturer of column' factor was changed to 'batch of column'. The manufacturer

was fixed in this method so it was more useful to test batch-to-batch variations. For the same reason, the 'manufacturer' factor was changed to 'batch' in the second test case. It was difficult to discover whether a different column was really from a different batch. The columns were thus selected with the difference between the serial numbers as large as possible.

The centrifugation time factor was removed from the factor choice because the evaluator uses another approach to centrifugation than the one used in RES. The expert system assumes a fixed centrifugation time, whereas the evaluator centrifuges until the solution is clear. Instead, two other sample preparation factors were added, shaking time and sonicating time, which the evaluator judged to be more important.

Although the evaluator changed the factor choice of the expert system considerably in one of the test cases, the general conclusion is that the factor choice is functioning properly and that the flexibility built into this module enhances its use. In fact, the changes made to the factor choice by the evaluator demonstrated the flexibility and usability of RES. The user can overrule the expert system whenever he thinks it is necessary. Similarly, the selection of design module performed to the satisfaction of the evaluator.

The evaluator suggests the inclusion of more flexibility in the calculation of results module. If an internal standard is applied there is no possibility to enter this information in the data file. In one of the test cases this was necessary for the calculation of the resolution between the internal standard and the compound of interest.

An additional test on the statistical significance of the main effects was also suggested. In the first test case, the method was not rugged to four factors with respect to the parameter concentration (Table 6). The evaluator tested the statistical significance of the main effect with respect to the standard deviations computed from the duplicates of every experiment. Through this test, the number of factors to which the method is not rugged is reduced to one.

The evaluator had problems with the interpretation of the results presented by the system. This was partly due to the lack of a good manual and

TABLE 6

Main effects for test case 2.1 of Evaluator 2

Level	Parameter	Component	Factor	Main effect (%)
upper	conc.	1	solvent	1.22 **
		2	data-handling	1.06
	retention	2	solvent	26.27
	peak height *	1	solvent	8.97
	(drift factors)	2	solvent	-7.65
lower	conc.	1	shake	1.29
		2	batch	-1.08
	retention	2	solvent	-18.52
	peak height *	1	solvent	-13.14
	(drift factors)	2	solvent	6.37

* With respect to the internal standard.

** The statistically significant main effect.

also partly due to the lack of explanation facilities in the specific version of the system. It is expected that both problems can be solved easily.

In general the evaluator judged the system to be useful. It produced good results and proved to give a valid set-up and interpretation of a ruggedness test. The evaluator adds a possible use of the expert system; the testing of methods developed at other laboratories on their ruggedness before using them in one's own laboratory.

DISCUSSION

In general, both evaluators agreed on the usefulness and the usability of RES. RES appeared to contain a valid strategy for ruggedness testing for an expert user as well as a normal user. The comments made by the evaluators contain a number of interesting suggestions concerning improvements to RES.

Sometimes, for the problems in category two, decisions have to be made about the desirability of including more than one strategy in the system. When strategies differ considerably a choice must be made of whether to include the different strategies or to include only one strategy. If more than one strategy is incorporated into the expert system, the user must be given a choice between the two or the expert system must decide between

the two. If the user must decide on which strategy to follow, some expert knowledge is required from the user. This is undesirable because a user will normally not have this knowledge available and will either have to consult an expert or make a decision himself without realizing the consequences. To let the expert system itself decide between two strategies is not always possible. Before such a decision can be incorporated into the expert system, the experts themselves must agree, which is sometimes not easily achieved. Therefore, it seems better to include only one strategy in the system. This will avoid confusion in the users because the system will always produce a consistent and reproducible reasoning process.

With respect to suggestions for the addition of parts to the system, the proposed extensions are in general very useful, especially where the two evaluators make the same suggestions. However, the expert system has been developed with a certain user in mind. If too many options are added to the system it may expect certain actions from the user that the user cannot perform. For instance, it was suggested that the statistical significance of the main effect be tested with a previously determined standard deviation of the method using an *F*-test with 95% confidence interval. This would probably reduce the number of main effects. If additional statistics are implemented, the user will have to enter a number of characteristics of the method, such as the reproducibility of the method. Such figures may not always be available. Another possibility is the addition of a repeatability test to RES in order to enable the user to measure the standard deviation of the method. Such an addition is the subject of current research [10].

The results of the test procedure described here give rise to a number of comments on the test strategy used. It can be concluded that the strategy was reasonably successful. Comments were made on most of the criteria, so a good estimate of the performance of the expert system could be made. Also, a number of additional comments were made, for instance, comments on the strategy employed by the expert system. One of the criteria that should have been tested—the explanation facilities of the expert system—has hardly been

tested in reality. This was mainly caused by practical constraints such as hardware that was not available at the evaluator's site. In retrospect, it is difficult to say which criteria should be added to the list so as to incorporate some of the additional comments made by the evaluators. This is probably caused by the very specific nature of the knowledge base.

The evaluation process has been a manual process in that it involved manual changes to parts of the knowledge base. Currently, automatic refinement of the knowledge base is under investigation, based on refinement strategies developed for rule-based systems [26,27]. Automatic refinement allows small changes to be made to the knowledge base when errors are found in the expert system's results. In particular, parts of the system that produce intermediate results, e.g. the factor choice, can be refined automatically. In these parts, the changes to the system's results made by the user are known in the system because they are entered by the user. An elaborate storage system and an evaluation of the various possible changes to the knowledge base are required to develop such an automatic refinement system. Recently, an application of such a system on a spectroscopic subject was reported [28].

CONCLUSIONS

In general the test procedure described here should be sufficient to establish the confidence of future users in the system. In the validation phase, the expert system has been tested on its performance in comparison with a human expert. It appeared to have a success rate of 80–90%. This is a nearly ideal situation because a 100% score is unrealistic and even undesirable. In the cases where there was disagreement between the expert and the expert system it was very difficult to decide which solution was better.

In the evaluation phase, the performance of the expert system has been estimated in practice. Both evaluators submitted a broad range of test cases to the system. They identified a number of recommendations that would improve the system, in their view, but they agreed upon the usefulness of

the system in practice. The system is probably more useful for a user inexperienced in ruggedness testing than for an experienced user. As the system was intended for inexperienced users, this was to be expected.

Evaluator 2 suggested that the system would also be useful to test the ruggedness of methods originating from other laboratories to be adopted by one's own laboratory. Evaluator 1 remarked that for him, it would be sensible to use only certain modules of the system and use their own strategy where it differed too much from the strategy embedded in RES.

REFERENCES

- 1 L. Buydens, A. Peeters and D.L. Massart, Knowledge representation for the selection of methods in analytical chemistry, *Chemometrics and Intelligent Laboratory Systems*, 5 (1988) 73–79.
- 2 D. Goulder, T. Blaffert, A. Blokland, L. Buydens, A. Chahabra, A. Cleland, N. Dunand, H. Hindriks, G. Kateman, H. van Leeuwen, D.L. Massart, M. Mulholland, G. Musch, P. Naish, A. Peeters, G. Postma, P. Schoenmakers, M. Desmet and B. Vandeginste, Expert Systems for Chemical Analysis (ESPRIT Project 1570), *Chromatographia*, 26 (1988) 237–243.
- 3 R. Hindriks, F. Maris, J. Vink, A. Peeters, M. DeSmet, D.L. Massart and L. Buydens, Expert system for the selection of initial high-performance liquid chromatographic conditions for the analysis of pharmaceuticals, *Journal of Chromatography*, 485 (1989) 255–265.
- 4 A. Peeters, L. Buydens, D.L. Massart and P.J. Schoenmakers, An expert system for the selection of criteria for selectivity optimization in high-pressure liquid chromatography, *Chromatographia*, 26 (1988) 101–109.
- 5 P.J. Schoenmakers, N. Dunand, A. Cleland, G. Musch and T. Blaffert, An expert system for the optimization of columns, operating conditions and instrumentation for high-pressure liquid chromatography, *Chromatographia*, 26 (1988) 37–44.
- 6 P.J. Schoenmakers and N. Dunand, Explanation and advice provided by an expert system for system optimization in high-performance liquid chromatography, *Journal of Chromatography*, 485 (1989) 219–236.
- 7 M. Mulholland, J.A. van Leeuwen and B. Vandeginste, An expert system for designing an intelligent spreadsheet for evaluation of precision of liquid chromatographic methods, *Analytica Chimica Acta*, 223 (1989) 183–192.
- 8 M. Mulholland, N. Dunand, A. Cleland, J.A. van Leeuwen and B.G.M. Vandeginste, Expert system for method validation in chromatography, *Journal of Chromatography*, 485 (1989) 283–296.

- 9 J.A. van Leeuwen, B.G.M. Vandeginste, G. Kateman, M. Mulholland and A. Cleland, An expert system for the choice of factors for a ruggedness test in liquid chromatography, *Analytica Chimica Acta*, 228 (1990) 145–153.
- 10 J.A. van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste and G. Kateman, Expert system for precision testing in validation of liquid chromatographic methods, *Analytica Chimica Acta*, 235 (1990) 27–40.
- 11 L.M.C. Buydens, J.A. van Leeuwen, M. Mulholland, B.G.M. Vandeginste and B.G.M. Kateman, An expert system for the validation of high performance liquid chromatographic methods, *Trends in Analytical Chemistry*, 9 (1990) 58–62.
- 12 J.A. van Leeuwen, B.G.M. Vandeginste, L. Buydens and G. Kateman, Expert systems in chemical analysis (ESCA), *Trends in Analytical Chemistry*, 9 (1990) 49–54.
- 13 J.A. van Leeuwen, B.G.M. Vandeginste, G.J. Postma and G. Kateman, An evaluation of expert system tools on their use in high-performance liquid chromatography, *Chemometrics and Intelligent Laboratory Systems*, 6 (1989) 239–252.
- 14 E.H. Shortliffe and R. Davis, *Newsletter 55*, ACM Special Interest Group on Artificial Intelligence (SIGART), December 1975.
- 15 J.A. van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, P.J. Schoenmakers and M. Mulholland, RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 1: The ruggedness test in HPLC method validation, *Chemometrics and Intelligent Laboratory Systems*, 10 (1991) 337–347.
- 16 J.A. van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, P.J. Schoenmakers and M. Mulholland, RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 2: The ruggedness expert system, *Chemometrics and Intelligent Laboratory Systems*, in press.
- 17 M.J. Cardone, Detection and determination of error in analytical methodology. Part 1: The method verification program, *Journal of the Association of Official Analytical Chemists*, 66 (1983) 1257–1282.
- 18 W.J. Youden and E.H. Steiner, *Statistical manual of the A.O.A.C.*, A.O.A.C., Washington, DC, 1975.
- 19 Guidelines for collaborative study procedure to validate characteristics of a method of analysis, *Journal of the Association of Official Analytical Chemists*, 72 (1989) 694–704.
- 20 M. Mulholland and J. Waterhouse, Development and evaluation of an automated procedure for the ruggedness testing of chromatographic conditions in high-performance liquid chromatography, *Journal of Chromatography*, 395 (1987) 539–551.
- 21 M. Mulholland and J. Waterhouse, Investigation of the limitations of saturated fractional factorial experimental designs, with confounding effects for an HPLC ruggedness test, *Chromatographia*, 25 (1988) 769–774.
- 22 G. Wernimont, Ruggedness evaluation of test procedures, *ASTM Standardization News*, March (1977) 13–16.
- 23 C.J.R. Green and M.M. Keyes, Verification and validation of expert systems, *Proc. of the 1987 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, New York, 1987, pp. 38–43.
- 24 J. Gaschnig, P. Klahr, H. Pople, E.H. Shortliffe and A. Terry, Evaluation of expert systems: issues and case studies, in B. Hayes-Roth, D.A. Waterman and D.B. Lenat (Editors), *Building Expert Systems*, Addison-Wesley, Reading, MA, 1983, pp. 241–280.
- 25 ANSI/IEEE std. 829–1983, *Software Test Documentation*, Wiley-Interscience, New York, 1983.
- 26 A.D. Preece, Verification of rule-based expert systems in wide domains, in N. Shadbolt (Editor), *Research and Development in Expert Systems VI*, Cambridge University Press, New York, Port Chester, Melbourne, Sydney, 1989, pp. 66–77.
- 27 A. Ginsberg, *Automatic Refinement of Expert System Knowledge Bases*, Pitman, London, 1988.
- 28 R. Wehrens, L. Buydens and G. Kateman, Validation and performance improvement of expert systems in analytical chemistry, *Chemometrics and Intelligent Laboratory Systems*, 12 (1991) in press.